Improving Prediction with Causal Probabilistic Variables

Ana Rita Nogueira[1,2(✉)], João Gama[1(✉)], and Carlos Abreu Ferreira[1(✉)]

[1] LIAAD - INESC TEC, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal
ana.r.nogueira@inesctec.pt, jgama@fep.up.pt, cgf@isep.ipp.pt
[2] Faculdade de Ciências da Universidade do Porto,
Rua do Campo Alegre 1021/1055, 4169-007 Porto, Portugal

Abstract. The application of feature engineering in classification problems has been commonly used as a means to increase the classification algorithms performance. There are already many methods for constructing features, based on the combination of attributes but, to the best of our knowledge, none of these methods takes into account a particular characteristic found in many problems: causality. In many observational data sets, causal relationships can be found between the variables, meaning that it is possible to extract those relations from the data and use them to create new features. The main goal of this paper is to propose a framework for the creation of new supposed causal probabilistic features, that encode the inferred causal relationships between the target and the other variables. In this case, an improvement in the performance was achieved when applied to the *Random Forest* algorithm.

Keywords: Causality · Causal discovery · Conditional probability · Feature engineering · Causal features

1 Introduction

In regular classification problems, a set of data, classified with a finite set of classes, is used as input so that the chosen classification algorithm can build a model, that represents the behaviour of the learning set. This classifier can have better or worse results, depending on the data and how the algorithm handles it.

Nevertheless, in many problems, applying only machine learning algorithms may not be the answer [4]. Instead, the use of feature engineering can be a way of improving the performance of these algorithms.

Feature engineering is a process by which new information is extracted from the available data, to create new features. These new features are related to the original variables, but also with the target variable, being a better representation of the knowledge embedded in the data, hence helping the algorithms achieve more accurate results [4]. These types of solutions are usually problem-related, being that one solution might work in one particular problem, but not

M. R. Berthold et al. (Eds.): IDA 2020, LNCS 12080, pp. 379–390, 2020.
https://doi.org/10.1007/978-3-030-44584-3_30

in the other. However, there is one particular characteristic common to many classification problems: causality. In observational data, there is the possibility of existing causal relationships between variables, especially in data related to medical problems (among others) [16,17]. This fact should be taken into consideration, for example when selecting or creating new features, since it can give clues to which variables are the most important to the problem.

By definition, causality, more specifically causal discovery, relates to the search of possible cause-effect relationships between variables [13]. The application of causal discovery in the various tasks of machine learning may be challenging, both at the level of the causal process or the sampling process to generate the observed data [9]. Despite this fact, this subject has been the focus of several researchers over the years, given the importance and the potential impact that the discovery of causal relationships between events can have in the problem-solving. In the words of Judea Pearl: *"while probabilities encode our beliefs about a static world, causality tells us whether and how probabilities change when the world changes, be it by intervention or by an act of imagination"* [20]. By discovering causal relationships, it is possible to uncover, not only correlations but also relations that explain how and why the variables behave the way they do.

In this paper, we propose a framework to create new features for discrete data sets (discrete features + discrete target) based on the causal relationships uncovered in the data. These attributes are created through the generation of a causal network, using a modified version of PC [21], and posterior probabilistic analysis of the relations a target variable and the variables considered as relevant. The relevant variables can be chosen by two different methods: parents and children of the target and Markov blanket [19].

This paper is organised as follows: Sect. 2 describes some important definitions. Section 3 describes the proposed framework and Sect. 4 the results obtained in the tests.

2 Background

In this section, we introduce some important notations that are used throughout the document.

2.1 PC

PC is a constraint-based algorithm and was proposed by Spirtes et al. [21]. This algorithm relies on the *faithfulness* assumption (*"If we have a joint probability distribution P of the random variables in some set V and a DAG G =(V,E), (G,P) satisfies the faithfulness condition if, G entails all and only conditional independencies in P"* [18]), meaning that all the independencies in a *DAG* (directed acyclic graph) need to respect the d-separation criterion [8].

This algorithm is divided into two phases. In the first phase, the algorithm starts with a fully connected undirected graph. It removes an edge if the two nodes are independent, *i.e.*, if there is a set of nodes adjacent to both variables in

which they are conditionally independent [12]. One of the most applied statistical independence tests is G^2, proposed by Spirtes et al. [21], and then used in non-causal Bayesian networks by Tsamardinos et al. [24].

In the second phase [12], the algorithm orients the edges by first searching for v-structures ($A \rightarrow B \leftarrow C$) and then by applying a set of rules, to create a completed partially directed acyclic graph ($CPDAG$), that is equivalent to the original one, where the faithfulness is respected.

2.2 Cochran-Mantel-Haenszel Test

The Cochran-Mantel-Haenszel test, [2] is an independence test that studies the influence of two variables on each other, and takes into account the possible influence of other variables on this dependence, *i.e.*, it searches for causal dependence [11].

There are two different versions of this test: the normal Cochran-Mantel-Haenszel test, which is used in $2 \times 2 \times K$ tables (being K the number of tables created), and the Generalised Cochran-Mantel-Haenszel tests, which is used in $I \times J \times K$ tables (being that I and J represent the number of categories in the studied variables, and K the number of layer categories [6]).

It is important to note that these type of contingency tables (three-way tables) are representations of the association between two variables if the influence of the other covariates is controlled.

Since many causal discovery algorithms (for discrete data) are used in data sets that are composed by a mixture of binary and non-binary discrete variables, the normal Cochran-Mantel-Haenszel test for $2 \times 2 \times K$ contingency tables is not enough. In such cases, the generalised version of this test can be applied instead (Generalised Cochran-Mantel-Haenszel test Eq. (1) [15]).

$$\begin{array}{ll} Q_{CMH} = G' Var\{G|H_0\}^{-1} G & G_h = B_h(n_h - m_h) \\ G = \sum_h G_h \quad Var\{G|H_0\} = \sum_h Var\{G_h|H_0\} & B_h = C_h \bigotimes R_h \end{array} \tag{1}$$

In the equations presented previously, B_h represents the product of Kronecker between C_h and R_h, Var the co-variance matrix, $(nh - mh)$ the difference between the observed and the expected, C_h and R_h the columns scores and row scores respectively, and H_0[1] the null hypothesis.

3 Framework

In many machine learning problems, the application of only classification algorithms might not be the answer to obtain satisfactory results [4]. The application

[1] *"For each of the separate levels of the co-variable set h = 1, 2, ..., q, the response variable is distributed at random with respect to the sub-populations,* i.e. *the data in the respective rows of the h_{th} table can be regarded as a successive set of simple random samples of sizes {Nhi.} from a fixed population corresponding to the marginal total distribution of the response variable {Nh.j}."* [15].

of feature engineering to the target data can be a way of improving such results. There are already several methods to improve the overall performance of an algorithm through the creation or modification of attributes, but, to the best of our knowledge, none of them explores the potential causal relationships between the target variable and the other variables.

The addition of these new inferred causal attributes may help improve the performance of classification algorithms, since they encode the relationship between the target and the other variables, thus feeding more information about the data set and its behaviour to the model. Moreover, these features may also aid in the generated models interpretability, since they encode the underlying relationships between the variables, thus being possible to explain more easily the decisions made by them.

In this section, we present a new framework to create new features using causal probabilities retrieve from a model that represents the causal associations between variables. This framework can be divided into four different phases:

1. Creation of the causal model (in this approach we suggest the usage of a modified version of PC);
2. Identification of the relevant variables. These variables are directly related to the target variable:
   - They are its parents and children;
   - They belong to its Markov blanket (*i.e.* parents, children and spouses).
3. Inference the probabilities associated with each pair {*target variable, associated variable*};
4. Creation of the new features using this probabilities. The number of features should be: $number\ of\ associated\ variables \times number\ of\ classes$.

In the first step, the framework starts by creating a full causal model, that represents the causal associations between all the variables. This is done through the application of a modified version of PC [21]. In this modified version, the state of the art independence test (usually χ^2 or G^2) is replaced by the Generalised Cochran-Mantel-Haenszel test presented in Sect. 2.2. This test has the advantage (over χ^2 and G^2) of adjusting for confounding factors [22].

It is important to note that, in some cases, PC can't direct every edge, hence it creates a CPDAG. In those cases, we apply a method to direct such edges. This method, proposed by Dor and Tarsi [5] searches recursively for possible ways to direct undirected edges.

In the second step, the framework selects the relevant variables. To select these attributes, we propose two different approaches: parents and children and Markov blanket.

In the parents and children (P-C) approach, as the name says, the variables selected are the ones that, in the causal graph, have an edge directed to the target (parents) or from the target (children).

In the Markov blanket (MB) approach, both the parents and children of the target are selected, as well as the nodes that have edges directed to the child nodes (also called spouse nodes). It is important to note that the most common way to select the variables that influence the target is through Markov Blanket

Table 1. Example of probabilities generated by the probability queries

		Attr		
		0	1	2
Target	0	0.63	0.53	0.13
	1	0.34	0.29	0.67
	2	0.14	0.25	0.56

(often used in causal feature selection methods [10]). However, several authors proposed to use only parents and children, as these variables can be considered to be the ones with the most influence in the target within its Markov blanket [1,3,23].

In the third step, the framework infers a set of probabilities that represent the influence of each relevant variable on the classes of the target: posterior probability distribution (Eq. (2)). In these probabilistic queries, the objective is to find what the influence that a evidence (particular values of the relevant variable) has on the value of the target [14]. This is performed for all the values in each variable and the resulting probability matrix is similar to Table 1.

$$P(Target = t|Attr = a) = \frac{occurrences_{t \cap a}}{ococcurrences_a} \tag{2}$$

Finally, in the fourth step, the new features are created and added to the data set. Each new feature represents the probability of the relevant variables influence on a specific class, *i.e.*, if we have, for example, a target variable with two classes ($\{0, 1\}$) and a relevant variable *Attr*, there will be created two new features representing the influence of *Attr* in each class (each instance of the feature represent the, influence the value of *Attr* in that instance on the class represented in that feature).

An overview of the framework can be seen in Fig. 1.

3.1 An Illustrative Example

To explain in more detail how this approach works, we will use as example a data set with 6 discrete variables (A, B, C, D, E and F), with 5000 instances. The values for variables A, B, C, D, and E can be $\{0, 1, 2\}$, while F can have the values $\{0, 1\}$. For this example, we will use variable **B** as the target.

As it was explained in *Step 1*, the approach starts by generating the full network with PC and Generalised Cochran Mantel Haenszel. The generated network can be seen in Fig. 2.

After the creation of the full network, the relevant variables are selected. The selected variables can be parents or children (P-C) of **B** ($\{A, E\}$) or the Markov blanket (MB) of **B** ($\{A, E, F\}$).

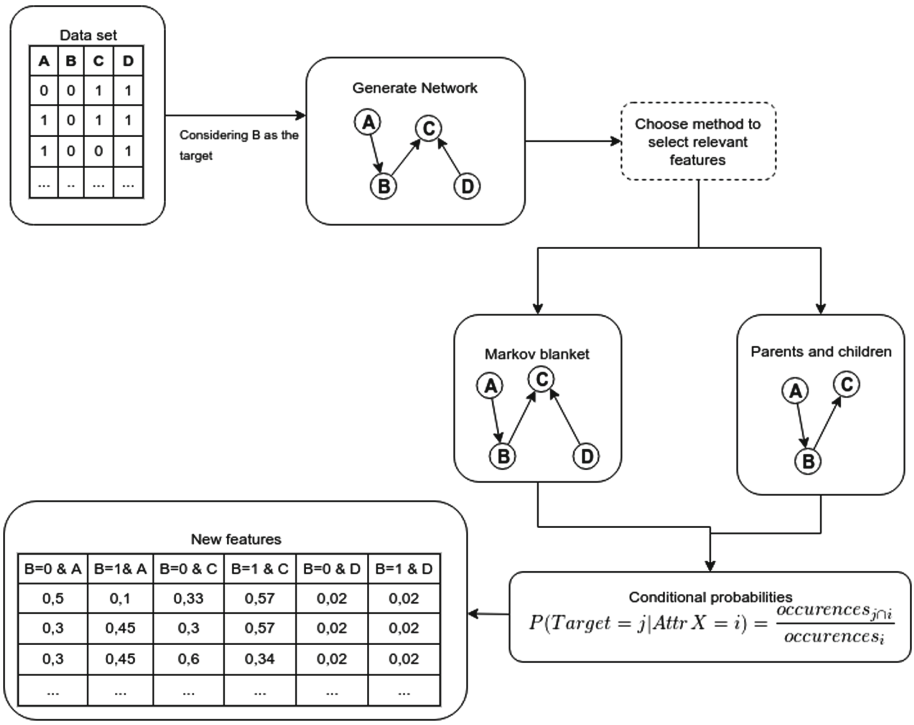

Fig. 1. Example of the operation of the proposed framework

In the third step of the framework generates inference probabilities for the chosen variables (Table 2). Taking A = 0 and B = 0 as an example, the probabilities are obtained for each one of the target values are obtained by by dividing the number of times both $A = 0$ and $B = 0$ occur, by the number of times A = 0 occurs, or in other words $P(B = 0|A = 0) = 0.86$.

These probabilities are then added to the global data set. The resulting data set is similar to Table 3. There is a difference between the number of new features created, since the number of generated features is equal to the product between the number of values in the target and the number of relevant variables. Since the MB approach selects more variables than the P-C approach, in theory, the number of generated features will be higher. So, in the case of P-C features we have 6 new features and in the case of MB we generate 9 new features.

4 Results and Discussion

To evaluate the proposed approaches and make a comparative study, the following configuration of experiments was designed: the performance of Random Forest, using the original data, as well has the versions generated by the two proposed approaches were compared.

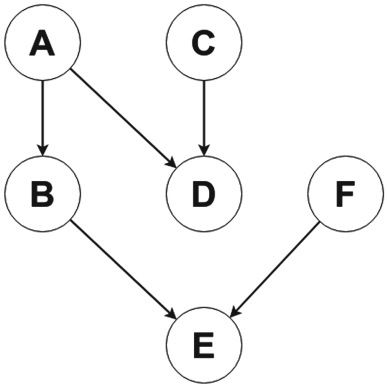

Fig. 2. Example: network generated

Table 2. Probabilities generated for the Markov blanket variables. In parents and children's case, the probabilities for F are not generated.

		A			E			F	
		0	1	2	0	1	2	0	1
Target	0	0.86	0.45	0.11	0.74	0.46	0.15	0.47	0.48
	1	0.03	0.22	0.09	0.08	0.11	0.16	0.11	0.12
	2	0.11	0.32	0.78	0.19	0.44	0.68	0.41	0.41

This comparative analysis was made through 10-fold cross validation, in several public data sets (Table 4). For each fold, the two approaches are applied to the train set and then the resulting conditional probabilities are used to create the new features for both the train and test set (this ensures that no information about the classes in the test set is added to the new features).

To choose the optimal parameters for the approaches presented in the following sections, a sensitivity analysis was performed. This analysis consisted of obtaining the error (*1 - accuracy*) for the presented data sets (by dividing them into 70% train, 30% test). In the case of PC this test was repeated for significance levels 1% and 5%. In these tests we concluded that the error of the algorithms in the data sets did not change much when the parameters were changed. For this reason, for all the data sets we select and present a significance level of 5%.

The performance of this algorithm was compared in terms of error rate (Table 5). This comparison was performed using the *No new features* as a reference. The classification algorithm performance, trained with causal features in each data set were compared to the reference using the Wilcoxon signed ranked-test. The sign +/− indicates that algorithm is significantly better/worse than the reference with a p-value of less than 0.05. Besides this, the algorithms are also compared in terms of average and geometric mean of the errors, average ranks, average error ratio, win/losses, significant win/losses (number of times

Table 3. Features generated with the probabilities for Markov blanket variables. In parents and children's case, the features related with F are not generated.

A	B	C	D	E	F	B = 2 A	B = 0 A	B = 1 A	B = 2 E	B = 0 E	B = 1 E	B = 2 & F	B = 0 & F	B = 1 & F
1	2	1	0	1	1	0.35	0.44	0.22	0.44	0.45	0.10	0.41	0.48	0.12
1	0	2	0	1	1	0.35	0.44	0.22	0.44	0.45	0.10	0.41	0.48	0.12
0	0	0	0	0	0	0.11	0.87	0.02	0.19	0.73	0.08	0.41	0.47	0.11
0	0	0	0	1	1	0.11	0.87	0.02	0.44	0.45	0.10	0.41	0.48	0.12
0	0	1	2	0	0	0.11	0.87	0.02	0.19	0.73	0.08	0.41	0.47	0.11

Table 4. Data set description

Data set	Number of examples	Number of attributes	Number of classes
breast cancer	286	10	0(70%) 1(30%)
cervical	858	16	0(94%) 1(6%)
corral	160	7	0(56%) 1(44%)
earthquake	10000	5	0(2%) 1(98%)
head injury	3121	11	0(92%) 1(8%)
lucas	2000	12	0(28%) 1(72%)
medpar	1495	9	0(66%) 1(34%)
mifem	1275	10	0(25%) 1(75%)
qualitative bankruptcy	250	7	0(43%) 1(57%)
respiratory	555	5	0(51%) 1(49%)
survey	10000	6	0(56%) 1(28%) 2(16%)
titanic	1316	4	0(62%) 1(38%)
xd6	973	10	0(67%) 1(33%)

that the reference was better or worse than the algorithm, using signed ranked-test) and the Wilcoxon signed ranked-test. For the Wilcoxon signed ranked-test we consider also a p-value of 0.05.

If we analyse Table 5, it is possible to see that, in general, *+Causal features P-C* (the addition of features representing the conditional probability of parents and children features on the target) has a better performance than *No new features*, since the value obtained in the Wilcoxon test is 0.0266 (less then the p-value of 0.05), which means that the difference between the performance is significant. This difference can also be seen in the values of the average and geometric ranks. More specifically, if we look at the average ranks, we can see that *+Causal features P-C* has lower ranks (in average) than *No new features* (1.436 against 2.538).

If we now compare the second approach proposed (*+Causal features MB*) with the reference, we can see that there is a positive difference in the results

Table 5. Error rates of Random Forest for classification with causal features

Data set	No new features	+Causal features P-C	+Causal features MB
breast cancer	28.6 ± 9.88	28.6 ± 7.49	28 ± 8.39
cervical	6.88 ± 1.51	6.65 ± 1.66	6.53 ± 1.49
corral	5.62 ± 5.47	+ 0.01 ± 0.10	+ 0.01 ± 0.10
earthquake	0.26 ± 0.14	0.20 ± 0.14	0.20 ± 0.14
head injury	7.08 ± 1.23	7.43 ± 0.83	7.05 ± 0.69
lucas	15.2 ± 2.02	14.5 ± 2.12	14.5 ± 2.12
medpar	32.70 ± 4.29	33.00 ± 3.91	34.10 ± 3.23
mifem	20.1 ± 4.28	20.00 ± 4.30	19.9 ± 3.63
qualitative bankruptcy	0.40 ± 1.26	0.01 ± 0.10	0.80 ± 2.53
respiratory	40.90 ± 6.79	40.20 ± 6.20	41.2 ± 6.90
survey	44.60 ± 2.26	44.4 ± 2.05	44.4 ± 2.05
titanic	21.4 ± 2.52	20.20 ± 2.19	20.5 ± 1.83
xd6	0.41 ± 0.72	0.10 ± 0.10	0.10 ± 0.10
Average Mean	17.242	16.562	16.715
Geometric Mean	7.161	2.889	4.039
Average Ranks	2.538	1.462	1.538
Average Error Ratio	1	0.764	0.914
Wicoxon test		0.0266	0.1465
Win/Losses		10/2	10/3
Significant win/losses		1/0	1/0

Table 6. AUC for Lucas data set

	AUC
No new features	0.877
+Causal features P-C	0.887
+Causal features MB	0.889

(although not significant). It is possible to see this difference, once again, in the average and geometric mean, as well as in the average rank (1.538).

In Table 6, it is possible to see the AUC values for the three analysed approaches, for lucas data set[2]. The results presented in this table were obtained by dividing this data set in train and test (70%/30%). The model scores were then obtained for the test data (with a 50% cutoff).

In this table it is possible to see that *+Causal features MB* has the highest area, meaning that, in the data set with the causal probabilistic features that represent the relations between the target and its Markov blanket, Random Forest can distinguish better the classes than with the data from the other approaches, thus having a better performance [7]. Although *+Causal features MB* was the

[2] http://www.causality.inf.ethz.ch/data/LUCAS.html.

best approach in terms of AUC, the other proposed approach *+Causal features P-C* also obtained an AUC higher than the reference.

Finally, from these results, we can conclude that there is evidence that applying causality to the creation of new features can have a positive impact on the classification algorithms performance.

5 Conclusion

The achievement of satisfactory results in a classification problem not only depends on the chosen classifier but also the data being processed. One possible way to improve the performance of classifiers is to apply feature engineering, or in other words, use the original data to infer new information, creating new attributes and altering others, to obtain more descriptive features. Furthermore, most of the proposed methodologies do not take into account the possible causal relationships in the data. This information can help to create more accurate models, since we are encoding in one variable, information about the interaction between variables, thus reinforcing their importance.

In this paper we proposed a framework that uses causal discovery to create new features based on posterior probabilistic analysis of the relations between a target variable and the variables considered as relevant, being these variables the parents and children of the Markov Blanket of the target.

In the experiments, we compared the approaches with the original data, using Random Forest in public data sets. From these results, we can conclude that there is evidence that the application of causality in the creation of new supposed probabilistic features may have a positive impact on the overall performance of the classification algorithm.

In the future, we intend to study the application of these techniques in other classifiers, as well as in the classification of mixed data (continuous and discrete variables).

Acknowledgments. This research was carried out in the context of the project FailStopper (DSAIPA/DS/0086/2018) and supported by the *Fundação para a Ciência e Tecnologia* (FCT), Portugal for the PhD Grant SFRH/BD/146197/2019.

References

1. Aliferis, C.F., Statnikov, A., Tsamardinos, I., Mani, S., Koutsoukos, X.D.: Local causal and Markov blanket induction for causal discovery and feature selection for classification part I: algorithms and empirical evaluation. J. Mach. Learn. Res. **11**(Jan), 171–234 (2010)
2. Birch, M.: The detection of partial association, I: the 2 × 2 case. J. Roy. Stat. Soc.: Ser. B (Methodol.) **26**(2), 313–324 (1964)
3. Bühlmann, P., Kalisch, M., Maathuis, M.H.: Variable selection in high-dimensional linear models: partially faithful distributions and the PC-simple algorithm. Biometrika **97**(2), 261–278 (2010)

4. Domingos, P.: A few useful things to know about machine learning. Commun. ACM **55**(10), 78–87 (2012)
5. Dor, D., Tarsi, M.: A simple algorithm to construct a consistent extension of a partially oriented graph. R-185, pp. 1–4, October 1992
6. Everitt, B.S.: The Analysis of Contingency Tables. CRC Press (1992)
7. Gama, J., Carvalho, A.C.P.d.L., Faceli, K., Lorena, A.C., Oliveira, M., et al.: Extração de conhecimento de dados: data mining. Sílabo (2015)
8. Geiger, D., Verma, T., Pearl, J.: d-separation: from theorems to algorithms. In: Machine Intelligence and Pattern Recognition, vol. 10, pp. 139–148. Elsevier (1990)
9. Glymour, C., Zhang, K., Spirtes, P.: Review of causal discovery methods based on graphical models. Front. Genet. **10**, 524 (2019). https://doi.org/10.3389/fgene.2019.00524
10. Guyon, I., Clopinet, C., Elisseeff, A., Aliferis, C.: Causal feature selection. Training **32**, 1–40 (2007)
11. Jin, Z., Li, J., Liu, L., Le, T.D., Sun, B., Wang, R.: Discovery of causal rules using partial association. In: Proceedings - IEEE International Conference on Data Mining, ICDM pp. 309–318 (2012)
12. Kalisch, M., Buehlmann, P.: Estimating high-dimensional directed acyclic graphs with the PC-algorithm. J. Mach. Learn. Res. **8**, 613–636 (2005)
13. Kleinberg, S.: Why: A Guide to Finding and Using Causes. O'Reilly Media Inc., Newton (2015)
14. Koller, D., Friedman, N.: Probabilistic Graphical Models: Principles and Techniques. MIT Press, Cambridge (2009)
15. Landis, J.R., Heyman, E.R., Koch, G.G.: Average partial association in three-way contingency tables: a review and discussion of alternative tests. Int. Stat. Rev./Revue Internationale de Statistique **46**(3), 237 (2006)
16. Listl, S., Jürges, H., Watt, R.G.: Causal inference from observational data. Commun. Dent. Oral Epidemiol. **44**(5), 409–15 (2016)
17. Martin, W.: Making valid causal inferences from observational data. Prev. Vet. Med. **113**(3), 281–297 (2014)
18. Neapolitan, R.E., et al.: Learning Bayesian Networks, vol. 38. Pearson Prentice Hall, Upper Saddle River (2004)
19. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Elsevier, Amsterdam (2014)
20. Pearl, J., Mackenzie, D.: The Book of Why: The New Science of Cause and Effect. Basic Books, New York (2018)
21. Spirtes, P., Glymour, C., Scheines, R.: Causation, Prediction, and Search, vol. 1, 2nd edn. The MIT Press, Cambridge (2001). https://ideas.repec.org/b/mtp/titles/0262194406.html
22. Tripepi, G., Jager, K.J., Dekker, F.W., Zoccali, C.: Stratification for confounding-part 1: the Mantel-Haenszel formula. Nephron Clin. Pract. **116**(4), c317–c321 (2010)
23. Tsamardinos, I., Aliferis, C.F., Statnikov, A.R., Statnikov, E.: Algorithms for large scale Markov blanket discovery. In: FLAIRS Conference, vol. 2, pp. 376–380 (2003)
24. Tsamardinos, I., Brown, L.E., Aliferis, C.F.: The max-min hill-climbing Bayesian network structure learning algorithm. Mach. Learn. **65**(1), 31–78 (2006). https://doi.org/10.1007/s10994-006-6889-7